Anesthesiology 2010; 112:1041–52

David S. Warner, M.D., Editor

Simulation-based Assessment in Anesthesiology

Requirements for Practical Implementation

John R. Boulet, Ph.D.,* David J. Murray, M.D.†

ABSTRACT

Simulations have taken a central role in the education and assessment of medical students, residents, and practicing physicians. The introduction of simulation-based assessments in anesthesiology, especially those used to establish various competencies, has demanded fairly rigorous studies concerning the psychometric properties of the scores. Most important, major efforts have been directed at identifying, and addressing, potential threats to the validity of simulation-based assessment scores. As a result, organizations that wish to incorporate simulation-based assessments into their evaluation practices can access information regarding effective test development practices, the selection of appropriate metrics, the minimization of measurement errors, and test score validation processes. The purpose of this article is to provide a broad overview of the use of simulation for measuring physician skills and competencies. For simulations used in anesthesiology, studies that describe advances in scenario development, the development of scoring rubrics, and the validation of assessment results are synthesized. Based on the summary of relevant research, psychometric requirements for practical implementation of simulation-based assessments in anesthesiology are forwarded. As technology expands, and simulation-based education and evaluation takes on a larger role in patient safety initiatives, the groundbreaking work conducted to date can serve as a model for those individuals and organizations that are responsible for developing, scoring, or validating simulation-based education and assessment programs in anesthesiology.

THE specific purpose of this article is to provide an overview of some of the issues that must be addressed to more fully embrace simulation-based methodology in the assessment of anesthesiologists. These assessments are formative (*e.g.*, education of residents), involving detailed participant feedback, or summative (*e.g.*, graduation requirement and board certification), with higher stakes consequences for those who participate. The following four general areas are highlighted: defining the pertinent skills and choosing relevant simulation tasks, establishing appropriate metrics, determining the sources of measurement error in test scores, and providing evidence to support the validity of test score inferences. For each of these areas, a general discussion is integrated with a brief review and synthesis of relevant anesthesia-related investigations. Because many of the logistic impediments have been addressed as part of recently established performance-based certification and licensure examinations,[1] and the specific challenges of integrating simulation into the existing anesthesia training curricula have been noted,[2] the discussion, in both general and specific to anesthesiology, will center on psychometric issues and not those associated with test administration logistics, physical test site specifications, or curriculum development. Knowing more about the specific psychometric challenges and potential solutions allows for further expansion of simulation-based assessment in anesthesiology. Before these challenges are outlined, a brief overview of the use of simulation, in both general and specific to anesthesiology, is presented.

Background

The need for assessments that encompass the skills required for specialty practice remains a priority of the Institute of Medicine. However, the Anesthesia Patient Safety Foundation, a pioneering organization in promoting safety, in responding to the report of the Institute of Medicine observed that these assessments "are not a simple matter and (defining and assessing competence in practice) will require considerable research."‡ Fortunately, previously conducted studies, both in anesthesiology and in other disciplines, have led to enhancements across the spectrum of simulation modalities, including advances in scenario design, the formulation and

* Associate Vice President, Research and Data Resources, Foundation for Advancement of International Medical Education and Research, Philadelphia, Pennsylvania. † Anesthetist-in-Chief, St. Louis Children's Hospital, St. Louis, Missouri, and Carol B and Jerome T Loeb Professor of Medicine, Director, Howard and Joyce Wood Simulation Center, Washington University School of Medicine, St. Louis, Missouri.

Received from Research and Data Resources, Foundation for Advancement of International Medical Education and Research, Philadelphia, Pennsylvania. Submitted for publication August 25, 2009. Accepted for publication November 3, 2009. Support was provided solely from institutional and/or departmental sources.

Address correspondence to Dr. Boulet: Associate Vice President, Research and Data Resources, Foundation for Advancement of International Medical Education and Research, 3624 Market Street, Philadelphia, Pennsylvania 19104. jboulet@faimer.org. This article may be accessed for personal use at no charge through the Journal Web site, www.anesthesiology.org.

‡ Stoelting RK: APSF Response to the IOM Report. 2009. Available at: http://apsf.org/about/pioneering_safety/. Accessed September 24, 2009.

utilization of sophisticated scoring algorithms, and the development of innovative methodologies to set performance standards. Moreover, considerable research has been undertaken with the express purpose of identifying potential threats to the validity of test score interpretations. Although there will certainly continue to be psychometric, and other, difficulties, past experience with performance-based assessments would suggest that most challenges can be addressed through focused research efforts.

Performance assessments in medicine have a long history.[3] Various simulation modalities have been used to assess student, resident, and practitioner competence as well as to identify curricular deficiencies.[4–9] Recently, based primarily on concerns related to physician competency and patient safety, summative assessments, including those targeting specific performance domains, have been incorporated into the process used to license and certify physicians.[10] In contrast to formative assessments, where the primary goal is to provide feedback to the individual concerning strengths and weaknesses, summative evaluation activities are meant to determine some endpoint status (*e.g.*, competent or not competent; ready to practice independently). Appropriately, these types of assessments, in addition to focusing on the evaluation of knowledge and clinical reasoning, have targeted other important competencies such as patient care, interpersonal and communication skills,[11] and teamwork.[12]

One of the main simulation modalities used to assess the clinical skills of physicians involves the employment of standardized patients, lay people who are trained to portray the mannerisms and model the complaints of real patients.[1,13–15] In developing these standardized patient-based simulations, especially those associated with certification and licensure, much was learned about examination design, test administration and logistics, quality assurance and, perhaps most important, psychometrics.[16,17] With respect to examination design, efforts were made to model simulation scenarios to specifically measure certain skills in a realistic way by choosing simulated patient complaints based on actual practice data.[18] Likewise, to ensure the precision of the scores, and any associated competency decisions,[19,20] both quantitative and qualitative performance-based quality assurance measures have been developed.[21] As testing practices evolve, and new simulation modalities emerge, they will need to be similarly scrutinized with respect to the reliability of the scores (or decisions made based on the scores), the validity of the inferences one can make based on the scores, and their overall fairness.

In anesthesiology, computer-based case management scenarios, task trainers, and mannequins have all been used as part of both formative and summative simulation-based assessments. From a general perspective, Seropian *et al.*[22,23] provided a detailed overview of the concepts and methodology of simulation in medicine. Issenberg *et al.*[24] summarize the benefits of simulation-based assessment and outline the use of simulation technology for healthcare professional skills training and assessment. Schwid[25] presents a synopsis of high fidelity mannequin-based simulations and the available technologies. Similarly, Cooper and Taqueti[26] provide a brief history of the development of mannequin simulators for clinical education and training. Cumin and Merry[27] review the current spectrum of anesthetic simulators and provide guidelines for their selection for specific tasks. Ziv *et al.*[28] provide an overview of credentialing and certification with simulation. There are even guidelines for those who want to include standardized patients in their anesthesia training programs.[29,30] Sinz[31] describes the history of anesthesiology's role in simulation and the efforts of the American Society of Anesthesiologists to promote simulation-based instruction. Scalese *et al.*[32] summarize the use of technology for skills training and competency assessment in medical education. Finally, going forward, Gaba[33] provides a future vision of simulation in healthcare. Taken together, these reviews, guidelines, and ideas delimit the potential uses of simulation technologies for the education and assessment of anesthesiologists.

Although much of the simulation work in anesthesiology has been limited to part-task (*e.g.*, airway management trainers) or full-body models (*i.e.*, electromechanical mannequins), the vast array of research conducted so far has advanced the entire field of performance assessment. In particular, several articles have specifically identified the numerous challenges and opportunities,[23,33–37] promoted the use of simulation-based assessment to identify individual skills deficiencies and associated curricular problems,[38–42] evaluated human factors and systems-level problems[43] and, as part of continuing medical education activities, advocated simulations for use as part of the assessment of anesthesiologists with lapsed skills.[44] As well, much of the research conducted so far has focused on the individual trainee or practitioner, including physicians in need of remediation[44] and those involved in self-directed lifelong learning activities.[24] Recently, the American Board of Anesthesiology outlined the four steps required for maintenance of certification in anesthesia. One step in the process involves the evaluation of practice ability; candidates can demonstrate this practice performance assessment and improvement at accredited simulation centers. Over their 10-yr Maintenance of Certification in Anesthesiology (MOCA) cycle, diplomates must complete two practice performance assessment and improvement activities. By including a step that requires practice performance assessment and improvement, albeit formative in nature, relatively infrequent, and not specifically associated with performance standards, the American Board of Anesthesiology recognizes the role of simulated environments in improving skill and expertise. Moreover, given the more recent emphasis on the education and evaluation of teams in high-acuity medical situations[45] and the assessment of interprofessional communication,[46] the high fidelity simulated environment offers the potential to assess many of the complex skills needed by specialists. Anesthesiologists can practice skills that improve their clinical and teamwork competencies, especially in preventing and managing critical events and maintaining their expertise in handling uncommon and rare events. To do this effectively, however, care must be taken to create simulation scenarios that yield meaningful scores.

Key Issues

Defining the Skills and Choosing the Appropriate Simulation Tasks

Although much has been written about the development of mannequin simulators and the design of educational programs,[9,22,24,47] the construction of quality simulation-based assessments continues to be a difficult task. Test developers must pay attention to the intended purpose of the test, the knowledge and skills to be evaluated and, for performance-based activities, the specific context for, and design of, the exercises.[48] To be effective, the assessment activities must also be targeted at the ability level of the examinee.[49] If the purpose of the assessment is not clear, then any ability measures that are gathered may be inaccurate. The choice of skills to be evaluated is usually guided by curricular information, competency guidelines,§ and the technical limitations of the chosen simulators.[35] Once these evaluation issues have been identified and synthesized, one is left with the task of specifying the simulation parameters. Most important among these is choosing the particular scenarios that offer the best opportunity to sample the knowledge and skills that one wishes to measure. For anesthesiology, and other specialties, one can access available data resources such as the National Hospital Ambulatory Medical Care Survey‖ to determine the most prevalent conditions and procedures. However, often the best opportunity to assess specific skill sets such as clinical decision making and communication is to select rare, or reasonably complex, events such as when air is entrained during an operation or when septic shock complicates the perioperative period. Based on existing performance-based certification and licensure examinations, an effective strategy has been to rely on both healthcare data resources, where available, and expert judgment.

With the rapid development of simulator technology, including full body mannequins and part-task trainers, the potential domain for assessment, both in terms of the skills being measured and the tasks that can be modeled, has greatly expanded.[27,50,51] For example, with the newer electromechanical mannequins, in addition to an inventory of preprogrammed scenarios, simpler and more intuitive programming interfaces allow faculty to model scenarios with realistic respiratory mechanics, carbon dioxide waveforms, and hemodynamic responses. For other simulation modalities such as standardized patients, it is often difficult, if not impossible, to measure procedural and management skills. Mannequins and part task trainers can be used to evaluate specific therapeutic skills (*e.g.*, airway management, venipuncture techniques, administering drugs, and cognitive steps in decision making) and, in combination with other simulation modalities, abilities related to resource management, professionalism, and teamwork.[52,53] Similar to the expansion of knowledge-based item-testing formats, the introduction of new simulation modalities provides an opportunity to measure various skills in different and more realistic ways, a change that should eventually yield more robust and defensible assessments.

Although the introduction of new simulation modalities will certainly expand the assessment domain, there are some limitations with current technologies, many of which have been acknowledged in the literature.[54] First, even with the most sophisticated high fidelity mannequins, some presentations and patient conditions cannot be modeled very well (*e.g.*, sweating, changes in skin color, and response to painful stimuli). As a result, there will still be a need to incorporate assessment activities that involve direct patient contact. Second, for electromechanical mannequins, the interrelationships between different physiologic variables can be imperfect, especially when attempting to simulate a patient with an unstable condition who then receives multiple pharmacologic interventions. If the simulator responds unpredictably to a given intervention (*e.g.*, coadministration of an anesthetic and an inotropic drug), whether this a function of canned programming or operator intervention, those being assessed may become confused and act in ways that are consistent with instructional feedback but inconsistent with intended patient care expectations. As a result, it will be difficult to have any confidence in the assessment results. Moreover, to the extent that those being assessed are continually queued by changes in monitored output, improperly scripted or modeled scenarios, or ones that are unnecessarily complex, will provide a poor milieu for evaluating ability.[35] Those charged with developing simulation-based assessments must balance the need to measure specific abilities with technological limitations of the simulators, recognizing that many conditions cannot be simulated with sufficient fidelity, potentially compromising stakeholder buy in.[55,56]

The scenario is the fundamental building block of most simulation-based assessments in anesthesiology. In general, the design and development procedures for a simulation-based assessment include the following: selecting competence domains that are amenable to a simulation environment, defining the expected skills that are needed to diagnose and manage the crisis, and designing a scenario that has the required skills embedded into the framework. In anesthesiology, the development of simulation scenarios, both computer- and mannequin-based, has been described in detail and typically involves a structured process to gather the insights and opinions of experts in the field.[57–59] The process of scenario development and selection can later be cross-referenced with curriculum, training, or certification expectations. For anesthesia specialty training, the Joint Council on In-Training Examinations, a committee of the American

§ Accreditation Council for Graduate Medical Education: ACGME Outcome Project. 2007. Available at: http://www.acgme.org/outcome/. Accessed November 25, 2009.

‖ National Center for Health Statistics: National Hospital Ambulatory Medical Care Survey (NHAMCS). 2007. Available at: http://www.cdc.gov/nchs/ahcd.htm/. Accessed November 25, 2009.

Society of Anesthesiologists and American Board of Anesthesiology, publishes a relatively detailed content outline that delineates the basic and clinical sciences areas (including anesthesia procedures, methods, and techniques) that a specialist must be knowledgeable about.# Scenarios for many of the content areas described in the outline can easily be simulated for both education and assessment. For example, the recognition and management of the side-effects of anesthetic drugs, respiratory depression, hypotension, anaphylaxis, cardiovascular events (arrhythmias, myocardial ischemia), surgical procedures (air entrainment, hemorrhage), and complications related to equipment failure can all be modeled in simulation scenarios. Overall, the simulated environment is an ideal setting to explore many of the conditions, side-effects, and complications that are listed as key content domains in the outline.

The practice domain of anesthesia is fairly well defined, and the majority of simulated scenarios tend to concentrate on the skill sets expected during a crisis. The rationale for selecting the crisis event as the content of a typical scenario is based on a number of considerations. First, a physician's failure to rapidly manage an acute-care event is often associated with an adverse patient outcome. In a critical patient care setting, or in a situation where unexpected anesthetic or surgical complications arise, the outcome may hinge on whether or not the anesthesiologist knows how to manage a crisis. Second, physicians, particularly residents, frequently manifest skill deficits in performing a logical, sequential, and timely patient evaluation. The "borderline" resident often struggles with setting priorities, managing time effectively, and recognizing when to call for help. In clinical practice, residents with serious skill deficits in these essential domains are often not recognized until multiple questionable judgments and skills deficits are observed in a crisis setting. The acute care simulation is useful in assessing a resident's skill in managing many of these common but difficult-to-evaluate skill sets.[59] Scenarios designed around acute care events typically include skills in setting priorities, generating hypotheses, processing knowledge, assigning probabilities, isolating important from unimportant information, integrating competing issues, acknowledging limits, and learning when to call for assistance.[60] Finally, critical events normally include a compressed timeline. A scenario designed to assess the dynamic, interrelated skills required to resolve a crisis quickly can tap numerous abilities, including communication, planning, and both inferential and deductive reasoning.

Specifying what needs to be assessed, both in practice and as part of educational activities is not necessarily complex. However, as some skills are quite difficult to measure (*i.e.*, teamwork),[61] and various practice situations (*i.e.*, ones involving multiple healthcare workers) are not easily modeled in the simulation environment, there remain difficult challenges. One of the most important of these, described in the next section, is the development of fair, reliable, and valid outcome measures.[37]

Developing Appropriate Metrics

If simulation-based activities are to be used for assessment-related activities, either formative or summative, appropriate metrics must be constructed. One needs to be reasonably certain that the scores, however gathered, reflect "true" ability. In anesthesiology, and other disciplines that use simulations as part of education and certification, creating these rubrics is certainly one of the main assessment challenges. Although much has been learned from the development of performance-based assessments of clinical skills that utilize standardized patients,[62] the adaptation of some of this knowledge to those types of simulations that would be appropriate for anesthesiology is not without difficulties. With this in mind, efforts to develop scoring metrics for high fidelity simulators are currently expanding at a rapid pace.[12,58,63–66]

Based on the literature related to simulation-based assessment in anesthesiology, two types of scores have predominated—explicit process and implicit process. Explicit process scores take the form of commonly used checklists or key actions. For a given simulation scenario, content experts (usually practitioners), often with the support of patient care guidelines, determine which actions, based on the presenting complaint, are important for the candidate (medical student, resident, practicing physician) to complete to properly manage the scenario.[63,64,67] For example, to manage intraoperative hypoxemia, an anesthesiologist should take a number of initial steps to correct hypoxia (100% O_2) as well as diagnose the cause (auscultation, evaluate lung compliance, carbon dioxide waveform, and others). These important management activities, when listed as checklist items or key actions, are the logical basis of the scoring rubric. While the weighting of checklist items may have little impact on the overall score, especially for more common clinical presentations where individual tasks are conditionally dependent,[68] this strategy may be appropriate for some acute care simulation events such malignant hyperthermia. Here, certain actions such as recognizing the condition and then administering dantrolene must be done to effectively manage the patient's condition. In essence, the shortening of checklists to essential key actions implicitly weights critical procedural or management tasks. Unfortunately, although checklists, or shorter key actions, have worked reasonably well and have provided modestly reproducible scores depending on the number of simulated scenarios,[58] they have been criticized for a number of reasons. First, checklists, while objective in terms of scoring, can be subjective in terms of construction.[69] Even if specific practice guidelines exist for given conditions informing what goes on the checklist, there can still be considerable debate as to which actions are important or necessary given the patient's condition. Without this consensus, one could

# American Board of Anesthesiology and American Society of Anesthesiologists: Content Outline: Joint Council on In-Training Examinations, American Board of Anesthesiology Examination Part I. 2009. Available at: http://www.asahq.org/publicationsAndServices/ITE_Part1ContentOutline_Revised2009.pdf. Accessed October 18, 2009.

certainly question the validity of the scenario scores. Second, the use of checklists often promotes rote behaviors such as using rapid-fire questioning techniques or performing many, some perhaps irrelevant, procedures. Third, and likely most germane for acute care simulations typical to anesthesia, it is difficult to take into account timing and sequencing when using checklists or key actions. Here, one can envision many scenarios where it is not only important what the physician does but also the order and timing. For example, in a scenario associated with a circuit leak, the participant who quickly recognizes and rapidly corrects the hypoventilation would more successfully avert a more serious prolonged period of hypoventilation leading to hypoxia. Although checklist-based timing has been used in some evaluations,[70,71] the order of actions, at least for explicit process-based scoring, is often ignored completely.

Implicit process scoring, where the entire performance is rated as a whole, can also be used in simulation-based assessments. In the physician community, there is often considerable reluctance to use rating scales, citing concerns regarding interrater reliability. However, based on the literature, holistic or global rating scales can be effective, psychometrically defensible tools for measuring certain constructs, especially those that are complex and multidimensional such as teamwork and communication.[65,72–74] In many situations, avoiding rating scales, simply because they involve expert judgment rather than the documentation of explicit actions, may not be advisable. With proper construction and effective rater training, they can be used to measure some important medical skills, including the nontechnical aspects of anesthesia practice. They also allow raters to take into account egregious actions and unnecessary patient management strategies (*e.g.*, performing a needle decompression of the left chest for a scenario requiring an endobronchial intubation), something that would be quite difficult to do with checklists or even key actions.[75] From a reliability perspective, even though two raters watching the same simulation encounter may not produce the exact same score, or scores, it is often possible to minimize this source of error. In addition, where systematic differences in rater stringency exist, score equating strategies can be used.[76] In many instances, especially those where multiple scenario assessments are used, one may actually prefer to sacrifice some measurement precision to achieve greater score validity.[17]

When developing rating scales (implicit process measures), evaluators often concentrate solely on the measurement rubric (*i.e.*, specification of the constructs that are going to be measured and deciding the number of score points for the scale), frequently ignoring any rater training or quality assurance regimes. Although physician raters are clearly content experts, this does not necessarily qualify them to be evaluators. Regardless of their clinical experience and capabilities, evaluators need to be trained to use implicit process measures. Training protocols can include specific rater exercises (*e.g.*, rating benchmarked performances), various quality assurance measures (*e.g.*, double rating a sample of examinee performances), and periodic refresher training. By developing a meaningful rubric and ensuring, through training, that the evaluators are providing ratings that accurately reflect examinee abilities, it is possible to minimize bias and produce more valid measures of performance.

The impetus to create physician-specific ability measures,[77,78] combined with advances in simulator technology, provides an opportunity to develop new, perhaps more meaningful, measurement rubrics. Many of the available electromechanical devices typically used in anesthesia training can generate machine-readable records of the physiologic responses of the mannequin as well as the treatments used during the simulation. Provided that the mannequin responds realistically and reproducibly to any intervention (*e.g.*, ventilation or drug therapy), and the timing of the actions can be demarked, then it should be possible to develop explicit performance measures that are based on patient (simulator) outcomes. For example, in a scenario associated with hypoventilation such as endotracheal tube obstruction or cuff leak, the changes in the mannequin's minute ventilation and carbon dioxide may serve as a reasonable measure of the participant's performance. While developing these types of scoring metrics will require some additional work in scenario design and construction, this approach may provide a more effective and timely method to assess performance and provide feedback to participants.

For anesthesiology-related simulation activities, studies have specifically reported the processes used to develop scoring rubrics[59,79] and, historically, the need to refine the existing rating systems.[80] Byrne and Greaves[81] provide a review of the assessment instruments used for anesthetic simulations, highlighting the need for better measurement tools. For checklists, key actions, and even holistic rating scales, attention has been paid to the construct being measured (*e.g.*, clinical decision making), the applicability of various simulation technologies for gathering performance measures and, where applicable, relevant practice guidelines. For many simulation-based assessment activities, expert panels, often as part of some structured Delphi technique, have developed the scoring rubrics.[64,67] Even though this strategy is appropriate, there is often little documented evidence regarding the qualifications of the panelists or the subsequent training of the raters. Although physicians are normally used as raters, and the rating is often accomplished *via* video review, establishing effective rater training programs is paramount, especially when the correctness of certain actions that are scored is open to some interpretation. Without this, any structured activity to develop scoring rubrics may still not yield meaningful scores.

For anesthesia simulations, "technical" skills such as airway management, drug administration, and placement of catheters have used checklists or key actions. In contrast, "nontechnical" skills such as communication, planning, teamwork and situation awareness, which play a key role in anesthesia,[82] have generally incorporated some form of holistic rating scale. Given the multidimensional nature of constructs such as communication skills, a more subjective rating methodology seems apropos. Various studies have looked at the relationships between scoring modalities, with some concluding that the relative ranking of participant abilities does not vary much whether holistic or analytic (checklist/key action) scores are used.[59,83] This finding will certainly depend on the type of simulation encounter used and the specific construct being measured. Based on the research conducted to date, the use of key action scores does seem to hold some advantages, at least for measuring procedural skills. First, at least for acute care scenarios, there is generally relatively little disagreement on what constitutes key actions. Second, they are relatively easy to score. Finally, if time stamps are used for critical actions, the sequencing of these actions can be captured. Overall, although much work has been has been conducted to develop meaningful rubrics for anesthesiology-based simulations, additional research aimed at specifically informing the construction and adoption of various measurement scales is certainly warranted.[81]

Assessing the Reliability of Test Scores

For simulation-based assessments, whether used for formative (*e.g.*, residency education) or summative (*e.g.*, certification or licensure) purposes, one needs to be reasonably confident that the scores are reliable. Compared with typical knowledge-based tests, there can be many other sources of measurement error in simulation-based assessments, including those associated with the rater.[20,84] If these sources, or facets, are not accounted for in the design, one can get an incomplete picture concerning the reliability of the scores. Where checklist or key actions are used to generate scores for a simulation scenario, internal consistency coefficients are typically calculated.[85] Although these coefficients can be presented as reliability indices, they provide only a measure of the consistency of examinee performance, across scoring items, within a scenario. For a multiscenario assessment, provided that care was taken in developing the scenario-specific performance measures, and specific skills are being assessed, one should be more concerned with the consistency of examinees' scores over encounters, and not within each individual one. Often, some measure of interrater reliability is also computed.[86,87] While scoring consistency between raters is certainly important, relying solely on a scenario-based measure of agreement is also incomplete. Even if two raters are somewhat inconsistent in their scoring, this may not necessarily lead to an unreliable total assessment score. To better understand the sources of measurement error in a multiscenario performance-based simulation assessment, generalizability (G) studies are often used.[88,89] These studies are conducted to specifically delimit the relative magnitude of various error sources and their associated interactions. Following the G-study, decision (D) studies can be undertaken to determine the optimal scoring design (*e.g.*, number of simulated encounters or number of raters per given encounter): that is, one that will provide sufficiently reliable scores given the purpose of the assessment.

Within the performance assessment literature, many studies have been conducted to estimate the impact of various designs on the reproducibility of the scores.[90] Although raters have been identified as a source of variability, their overall impact on reliability, given proper training and well-specified rubrics, tends to be minimal, often being far outweighed by task sampling variance. Essentially, given the content specificity of certain simulation tasks, especially those that are designed to assess technical skills, examinees may perform inconsistently from one task (or scenario) to the next. For example, based on previous experience and training, a participant may effectively recognize and treat anaphylaxis, yet fail to diagnose or effectively manage myocardial ischemia. As a result, if there are few tasks, and one is attempting to measure overall clinical ability, the reliability of the assessment score can be poor. To think of this concept another way, if we are trying to measure skills in patient management, for example, more performance samples (simulated scenarios) will mitigate the overall impact of content specificity, thus yielding a more precise overall ability measure. In general, for these types of performance-based assessments, issues regarding inadequate score reliability can be best addressed by increasing the number of simulated tasks rather than increasing the number of raters per given encounter or simulated scenario. As well, to minimize any rater effects, it is usually most effective to use as many different raters as possible for any given examinee (*e.g.*, a different rater for every task or scenario).[17]

In anesthesiology, there are some unique challenges associated with the development of simulation-based assessments, especially those where fairly reliable estimates of ability are required. Unlike many performance-based assessments in clinical medicine, where fairly generic skills are being measured (*e.g.*, history taking), the management of patients by anesthesiologists can be very task specific. Where this is true, and one wants to measure skills related to patient management, it could take many more encounters to achieve a reproducible measure of ability. Fortunately, many important events in anesthesia practice, including a large number that can be effectively modeled in a simulated environment, require fairly rapid interventions. Unlike typical standardized patient-based cases, which usually last from 10 to 20 min, acute care scenarios can easily be modeled to take place in a 5-min period. Given that the simulation scenarios can be relatively short, it is possible to include more of them in an assessment. Moreover, for nontechnical skills such as communication, task specificity would likely not be as great. Here, one would not expect that an individual's ability to communicate with the patient, or other healthcare provider, would vary much as a function of type of simulated event. Therefore, fewer behavioral samples (scenarios) would be needed to yield a reliable total score.

Given the amount of work that has been conducted to develop scoring rubrics for anesthesia-related simulation assessments,[64] it is not surprising that efforts have been directed at disentangling various sources of measurement error in the scores. As is common in medicine, the initial focus of many psychometric investigations rests with establishing interrater reliability. As noted previously, while it is important that two or more raters viewing the same performance are reasonably consistent in their evaluations, this is only one facet to take into account.[16] Interestingly, some investigations provide evidence to support rater consistency[74,91–94] while others do not.[12,95] While the exact cause for these disparate findings is hard to pinpoint, it likely rests with differences in the skills being assessed, the assessment mode (*e.g.*, live *vs.* videotape review), the choice of scoring rubrics, and whether specific rater training protocols are used.[16] All these factors could have some measurable impact on rater consistency. Recently, there has been a general recognition that obtaining reasonably precise measures of ability requires multiple scenarios or tasks sampled over a relatively broad content domain.[96] However, some behavioral attributes such as communication and teamwork may be less dependent on content knowledge, thus requiring fewer performance samples. In an investigation of the psychometric properties of a simulation-based assessment of anesthesiologists, Weller *et al.*[97] reported that 12–15 scenarios were needed to reliably rank trainees on their ability to manage simulated crises. Several studies of anesthesia residents and anesthesiologists have incorporated multistation assessments, reporting reasonable interstation reliabilities for evaluations that incorporate 8–12 scenarios.[58,59,98]

Providing Evidence to Support the Validity of Test Score Inferences

Validity relates to the inferences that we want to make based on the assessment scores. Inspecting the simulation literature, in general, and the research related to performance-based examinations, in particular, there are numerous potential sources of evidence to support the validity of test scores and the associated inferences made based on these scores.[99–101] However, it should be noted that the validation process is never complete and that scores may be valid for one purpose and not for another. Additional evidence to support the intended test score inferences can always be gathered.

For performance-based assessments, there has been a heavy emphasis on content related issues.[102,103] To support the content validity of the assessment, simulated scenarios are often modeled and scripted based on actual practice characteristics, including the types of patients that are normally seen in particular settings. With respect to rubrics, special care is usually taken to define the specific skill sets and to develop measures, often from an evidence-based perspective, that adequately reflect them. Finally, the encounters are typically modeled in realistic ways, using the same equipment that would be found in a real operating theater or other venue. All of these strategies, including feedback from stakeholders regarding the verisimilitude of the simulated scenarios,[104] will help support the content validity of the test scores.

If a simulation-based assessment is designed to measure certain skills, then it is imperative that evidence be procured to support this claim.[105] Various strategies can be used to accomplish this goal. If several skills are being measured, then one could postulate relationships among them. For example, if the simulation is designed to measure both procedural and communication skills, then one would expect that the scores for these two domains should be somewhat, albeit imperfectly, related. Likewise, if external measures are available (*e.g.*, knowledge-based in-training examination scores) one might postulate both strong and weak relationships between the simulation assessment scores and these criterion measures. Often, the criterion measure is some measure of clinical experience. Here, one would normally expect that individuals with greater expertise (*e.g.*, more advanced training or board certification) and having proper training will perform better on the simulation tasks.[58,71,94,106,107] If this is not the case, then one could question whether valid inferences can be made based on the assessment scores. Overall, to the extent that postulated relationships, both internal and external, substantiate the hypothesized relationships, support for the validity of test score interpretations can be gathered.

Unlike the more common formative simulation-based assessments, the purpose of some simulation-based evaluations is to ensure the public that the individual who passes the examination is fit to practice, either independently or under supervision. Here, it is imperative that score-based decisions are valid and reproducible. To accomplish this, a variety of standard setting techniques are available, some of which have been applied for acute care mannequin-based assessments.[108] As part of a structured process, subject-matter experts make judgments, either based on the score scale or some sampling of performances, concerning minimal competency as it relates to the particular simulation task. Using various statistical procedures, these judgments are then used to derive an appropriate cut-score, the point on the score scale that separates those who are minimally adequate (or some other definition) from those who are not. Unfortunately, while defensible cut-scores can be established for performance-based assessments, procuring evidence to support the validity of the associated decisions can be complicated.[109] Although performance on the simulation-based assessment may be indicative of future aptitude or competence, and there are some longitudinal studies that support this for certain skills,[110] the predictive relationships may be weak and difficult to measure.[111] From a research perspective, only those who "pass" the initial assessment can be studied; individuals who do not demonstrate competence are generally not allowed to practice, effectively attenuating any relationships between the assessment scores and any future outcome measure. Nevertheless, the introduction of simulation-based assessment, if done correctly, can provide the public with greater assurance that practitioners are qualified.[44,112] Also, if the consequential impact of other previously implemented assessments is a guide, this will ultimately lead to a growth in simulation-based educational programs, a change that will likely lead to greater patient safety.[113–115]

Although the simulation-based training of anesthesiologists has taken place for some time, and there were early calls for establishing the efficacy of this training,[116] more rigorous studies aimed at establishing the validity of assessment scores have come only more recently. From a content validity perspective, simulation scenarios have been modeled on real patient events and have included scoring rubrics that are keyed to practice-based guidelines. Moreover, in addition to specific patient management tasks, simulation scenarios have been developed to specifically target nontechnical skills such as communication, teamwork, and clinical decision making.[12,117,118] Although not generally considered strong evidence for validity, several studies have provided data summarizing the opinions of those being assessed. Based on various simulation modalities, and a host of clinical presentations, most studies indicate that participants thought that the exercises were realistic and pedagogically useful with respect to clinical training and competency assessment.[119,120] Berkenstadt *et al.*,[91] based on simulations incorporated in the Israeli Board Examination in anesthesiology, reported that those exposed to this form of assessment preferred it to the traditional oral examination. Given that trainees need to demonstrate specific skills, Savoldelli *et al.*[92] also supported the use of simulations as an adjunct to the oral examination for senior anesthesia residents. As simulation technology expands, the breadth of clinical scenarios that can be modeled will certainly increase, providing additional opportunities to gather content validity evidence.

Other sources of validity evidence have been reported throughout the literature. If the scoring systems are appropriate, and actually measure the intended construct, or constructs, then one would expect that those individuals with more training and experience would perform better. Likewise, given the effects of experiential education, especially if it involves repetitive practice and appropriate feedback, those being trained with simulators should show some performance gains over time or with additional training.[83,121] Going back almost 20 yr, a number of studies involving the assessment of medical students, residents, and practicing anesthesiologists have demonstrated this finding.[64,122] Moreover, individuals participating in simulator training have been able to retain their skills over time.[123] In addition to the evidence that supports the discriminant validity of the simulation-based exercises,[98,124] some studies have looked at the relationships between simulation scores and other measures of performance such as written tests of knowledge, course grades, and various nonsimulation-based resident evaluations.[125] From a criterion-related validity perspective, some studies have shown a moderate relationship between simulation performance and knowledge. Schwid *et al.*[94] reported positive correlations between simulation scores and both faculty evaluations and American Board of Anesthesiology written in-training examination scores. While other studies have shown little relationship between simulation scores and other evaluations,[126,127] this may be a function of differences in the constructs that are being measured. Investigators are quick to acknowledge that knowing what to do, which can be measured in many different ways, is somewhat different from showing what you can actually do, either in a real or simulated environment. As an example, one could envision an anesthesiology resident who performs well on in-training assessments, indicating knowledge of what to do in certain situations, but cannot effectively use this knowledge in managing a real or simulated event. To explain some simulator-criterion relationships, or lack thereof, one must not forget that to effectively use simulators as assessments devices, those individuals being evaluated must have some familiarity with the devices.[119] Often, the orientation process is insufficient. As a result, one might expect only moderate associations between simulator performance and other, perhaps marginally related, ability measures.

From a validity perspective, the strongest evidence lies with the establishment of a link between simulator performance and practice with real patients. To date, there have been relatively few studies that have shown a significant impact of simulator training from a patient outcome perspective. Unlike many other disciplines, there has, however, been some excellent work to show that skills acquired in the simulation environment transfer to "real world" situations. For example, Weller *et al.*[128] used simulation scenarios to investigate the impact of trained assistance on error rates in anesthesia and concluded that a simulation-based model can provide rigorous evidence on safety interventions. For resuscitation skills, Domuracki *et al.*[129] found that learning on the simulator, provided there was appropriate feedback, transferred to clinical practice. Kuduvalli *et al.*[130] also reported long-term retention and transferability of acquired skills into subsequent clinical practice. Unfortunately, these findings were only based on questionnaire responses. Although it can be considered as indirect validity evidence, Blum *et al.*[131] reported that training courses, often using simulation, can make faculty staff eligible for malpractice premium reductions. Even with these sources of validity evidence, there is still a need to continue to address the long-term effects of experiential learning on the retention of knowledge and acquired skills. More important, while extremely difficult to do, establishing a causal link between simulator performance and actual patient outcomes is essential.[132]

Conclusion

The expansion of simulation models, including those using mannequins and part-task trainers, will lead to many more opportunities to model real-life events. This, in turn, will demand additional investigations to support the use of the resulting assessment scores, either for educational activities (*e.g.*, for providing feedback to anesthesia trainees) or, more importantly, for higher stakes decisions concerning provider competency, including those associated with licensure and maintenance of certification activities.

From a development perspective, defining the skills and choosing appropriate simulation tasks is paramount. Simulation scenarios constructed to measure teamwork and communication skills may not be suited to measure procedural skills. Ideally, simulation scenarios should be modeled on real events and constructed in such a way as to provide the best milieu to evaluate the skills of interest, whether technical or nontechnical. To accomplish this goal, expert opinion, combined with relevant practice guidelines (if they exist), are the key elements. If the development of the simulation scenarios is not sufficiently rigorous, the assessment scores, regardless of the purpose of the evaluation, or who is being evaluated, may not have much meaning.

Once the content area, or areas, has been identified, developing appropriate metrics for simulation-based assessment activities is paramount. Whether one is providing residents with feedback, or assessing competence as part of certification or licensure activities, scores are needed. In general, the choice of metrics will depend on what is being measured. For technical skills (*e.g.*, airway management), it is usually possible to identify observable key actions and develop analytic measurement tools. For nontechnical skills, such as communication and teamwork, rating scales are usually more appropriate. Regardless of the metric that is chosen, care must be taken to identify the elements, or behaviors, that anchor the score scale. For key actions, the raters must know when, and when not, to give credit. For holistic, or global, rating tools, the raters must be clear about the construct being measured (*e.g.*, teamwork) and how someone who is more able in this domain differs from someone who is less able.

For most simulation-based assessments, estimating the reliability scores is not that difficult. Addressing the various sources of measurement can, however, be quite challenging. For situations where the scores are being used for higher stakes purposes, aggregate scores from multiple scenarios will generally be needed to obtain a sufficiently reliable estimate of ability. One should think of the scenarios as vehicles to measure the skills—more scenarios, or testing time, will, in general, yield more reliable estimates of ability. Although one should also be concerned with potential rater effects (*i.e.*, interrater inconsistency), rater training, combined with various quality assurance activities, will help minimize this potential source of error. Those charged with implementing simulation-based assessments in anesthesiology must identify the various sources of measurement error and use this information, where possible, to modify the structure of their evaluations. If an individual's score is not a sufficiently precise measure of his/her ability, actions based on this score (*e.g.*, ranking within the class, the provision of feedback, and certification decisions) could be misleading or erroneous.

For simulation-based assessments, providing evidence to support the validity of test score inferences is essential. Even for lower stakes formative assessment activities, including practice performance assessment and improvement initiatives and maintenance of certification in anesthesiology-related activities, one needs to know that the scores are reasonably accurate reflections of the skills that are purportedly being evaluated. In anesthesiology, much work has been conducted to gather evidence to support the validity of simulation-based assessment scores. These efforts should ultimately lead to better, more defensible, assessments, ones that can be used to identify individual strengths and weaknesses, including competency deficits. Going forward, outcomes-based research centering on quantifying the paths between simulation-based assessment, skills acquisition (and decay), and patient safety are essential. Without additional construct validity evidence, the utility of simulation-based assessment, at least for higher stakes applications such as board certification and licensure (or maintenance of licensure), is likely to continue to be questioned.

Anesthesiology as a specialty has made numerous prescient commitments to safer patient care. The adoption of simulation by anesthesiologists was eventually recognized as an assessment modality that can overcome many of the inherent patient risks involved in specialty training. A physician's advanced diagnostic and therapeutic management skills, and the ability to integrate knowledge, apply clinical judgment, communicate, and work within a team, can all be assessed during a high fidelity simulation. These types of performance assessments, when constructed with care and appropriately validated, are considered an essential element in elevating practice standards and, ultimately, in improving the safety of anesthesia practice. Through years of research, the breadth of simulation activities in anesthesiology has widened, with model-based training and assessment, albeit currently limited in scope, now accepted as one of the steps to maintain certification in the profession. By adopting simulation-based training and assessment, and actively addressing many of the challenges associated with developing psychometrically sound evaluations, the specialty has recognized the need for professional skill development, continuing on a path demonstrating a long-term commitment to patient care.

References

1. Scoles PV, Hawkins RE, LaDuca A: Assessment of clinical skills in medical practice. J Contin Educ Health Prof 2003; 23:182–90
2. Chin C, Arrica M, Bertolizio G, Ingelmo P: Simulation training in pediatric anesthesia. Minerva Anestesiol 2009; March 31 [Epub ahead of print]
3. Hubbard JP, Levit EJ, Schumacher CF, Schnabel TG, Jr: An objective evaluation of clinical competence. New technics used by the National Board of Medical Examiners. N Engl J Med 1965; 272:1321–8
4. Epstein RM: Assessment in medical education. N Engl J Med 2007; 356:387–96
5. Grenvik A, Schaefer JJ III, DeVita MA, Rogers P: New aspects on critical care medicine training. Curr Opin Crit Care 2004; 10:233–7
6. Berkenstadt H, Erez D, Munz Y, Simon D, Ziv A: Training and assessment of trauma management: The role of simulation-based medical education. Anesthesiol Clin 2007; 25:65–74
7. Bradley P: The history of simulation in medical education and possible future directions. Med Educ 2006; 40:254–62
8. Srinivasan M, Hwang JC, West D, Yellowlees PM: Assess-

ment of clinical skills using simulator technologies. Acad Psychiatry 2006; 30:505–15
9. Cooper JB, Taqueti VR: A brief history of the development of mannequin simulators for clinical education and training. Qual Saf Health Care 2006; 13(suppl 1):i11–8
10. Boulet JR, Smee SM, Dillon GF, Gimpel JR: The use of standardized patient assessments for certification and licensure decisions. Simul Healthc 2009; 4:35–42
11. Boulet JR, Ben David MF, Ziv A, Burdick WP, Curtis M, Peitzman S, Gary NE: Using standardized patients to assess the interpersonal skills of physicians. Acad Med 1998; 73:S94–6
12. Morgan PJ, Pittini R, Regehr G, Marrs C, Haley MF: Evaluating teamwork in a simulated obstetric environment. ANESTHESIOLOGY 2007; 106:907–15
13. Melnick DE, Dillon GF, Swanson DB: Medical licensing examinations in the United States. J Dent Educ 2002; 66:595–9
14. Dillon GF, Clyman SG, Clauser BE, Margolis MJ: The introduction of computer-based case simulations into the United States Medical Licensing Examination. Acad Med 2002; 77:S94–6
15. Gimpel JR, Boulet JR, Errichetti AM: Evaluating the clinical skills of osteopathic medical students. J Am Osteopath Assoc 2003; 103:267–79
16. Edler AA: The use of simulation education in competency assessment: More questions than answers (letter). ANESTHESIOLOGY 2008; 108:167
17. Boulet JR, Swanson DB: Psychometric challenges of using simulations for high-stakes assessment, Simulations in Critical Care Education and Beyond. Edited by Dunn WF. Des Plains, Society of Critical Care Medicine, 2004, pp 119–30
18. Boulet JR, Gimpel JR, Errichetti AM, Meoli FG: Using National Medical Care Survey data to validate examination content on a performance-based clinical skills assessment for osteopathic physicians. J Am Osteopath Assoc 2003; 103:225–31
19. Cook DA, Beckman TJ: Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med 2006; 119:166.e7–16
20. Downing SM: Reliability: On the reproducibility of assessment data. Med Educ 2004; 38:1006–12
21. Boulet JR, McKinley DW, Whelan GP, Hambleton RK: Quality assurance methods for performance-based assessments. Adv Health Sci Educ Theory Pract 2003; 8:27–47
22. Seropian MA: General concepts in full scale simulation: Getting started. Anesth Analg 2003; 97:1695–705
23. Seropian MA, Brown K, Gavilanes JS, Driggers B: Simulation: Not just a manikin. J Nurs Educ 2004; 43:164–9
24. Issenberg SB, McGaghie WC, Hart IR, Mayer JW, Felner JM, Petrusa ER, Waugh RA, Brown DD, Safford RR, Gessner IH, Gordon DL, Ewy GA: Simulation technology for health care professional skills training and assessment. JAMA 1999; 282:861–6
25. Schwid HA: Anesthesia simulators—technology and applications. Isr Med Assoc J 2000; 2:949–53
26. Cooper JB, Taqueti VR: A brief history of the development of mannequin simulators for clinical education and training. Postgrad Med J 2008; 84:563–70
27. Cumin D, Merry AF: Simulators for use in anaesthesia. Anaesthesia 2007; 62:151–62
28. Ziv A, Rubin O, Sidi A, Berkenstadt H: Credentialing and certifying with simulation. Anesthesiol Clin 2007; 25: 261–9
29. Levine AI, Swartz MH: Standardized patients: The "other" simulation. J Crit Care 2008; 23:179–84
30. Waisel DB, Simon R, Truog RD, Baboolal H, Raemer DB: Anesthesiologist management of perioperative do-not-resuscitate orders: A simulation-based experiment. Simul Healthc 2009; 4:70–6
31. Sinz EH: Anesthesiology national CME program and ASA activities in simulation. Anesthesiol Clin 2007; 25:209–23
32. Scalese RJ, Obeso VT, Issenberg SB: Simulation technology for skills training and competency assessment in medical education. J Gen Intern Med 2008; 23(suppl 1):46–9
33. Gaba DM: The future vision of simulation in healthcare. Simul Healthc 2007; 2:126–35
34. Kapur PA, Steadman RH: Patient simulator competency testing: Ready for takeoff? Anesth Analg 1998; 86:1157–9
35. McIntosh CA: Lake Wobegon for anesthesia… where everyone is above average except those who aren't: Variability in the management of simulated intraoperative critical incidents. Anesth Analg 2009; 108:6–9
36. Glavin RJ, Gaba DM: Challenges and opportunities in simulation and assessment. Simul Healthc 2008; 3:69–71
37. McGaghie WC, Issenberg SB, Gordon DL, Petrusa ER: Assessment instruments used during anaesthetic simulation. Br J Anaesth 2001; 87:647–8
38. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J: Identification of gaps in the achievement of undergraduate anesthesia educational objectives using high-fidelity patient simulation. Anesth Analg 2003; 97:1690–4
39. Hunt EA, Vera K, Diener-West M, Haggerty JA, Nelson KL, Shaffner DH, Pronovost PJ: Delays and errors in cardiopulmonary resuscitation and defibrillation by pediatric residents during simulated cardiopulmonary arrests. Resuscitation 2009; 80:819–25
40. Rose SH, Long TR, Elliott BA, Brown MJ: A historical perspective on resident evaluation, the Accreditation Council for Graduate Medical Education Outcome Project and Accreditation Council for Graduate Medical Education duty hour requirement. Anesth Analg 2009; 109: 190–3
41. Shilkofski NA, Nelson KL, Hunt EA: Recognition and treatment of unstable supraventricular tachycardia by pediatric residents in a simulation scenario. Simul Healthc 2008; 3:4–9
42. Daniels K, Lipman S, Harney K, Arafeh J, Druzin M: Use of simulation based team training for obstetric crises in resident education. Simul Healthc 2008; 3:154–60
43. Small SD: Simulation applications for human factors and systems evaluation. Anesthesiol Clin 2007; 25:237–59
44. Rosenblatt MA, Abrams KJ: The use of a human patient simulator in the evaluation of and development of a remedial prescription for an anesthesiologist with lapsed medical skills. Anesth Analg 2002; 94:149–53
45. Sundar E, Sundar S, Pawlowski J, Blum R, Feinstein D, Pratt S: Crew resource management and team training. Anesthesiol Clin 2007; 25:283–300
46. Matveevskii AS, Gravenstein N: Role of simulators, educational programs, and nontechnical skills in anesthesia resident selection, education, and competency assessment. J Crit Care 2008; 23:167–72
47. Dawson S: Procedural simulation: A primer. Radiology 2006; 241:17–25
48. Kneebone RL, Nestel D, Vincent C, Darzi A: Complexity, risk and simulation in learning procedural skills. Med Educ 2007; 41:808–14
49. McLaughlin SA, Doezema D, Sklar DP: Human simulation in emergency medicine training: A model curriculum. Acad Emerg Med 2002; 9:1310–8
50. Bond WF, Lammers RL, Spillane LL, Smith-Coggins R, Fernandez R, Reznek MA, Vozenilek JA, Gordon JA: The use of simulation in emergency medicine: A research agenda. Acad Emerg Med 2007; 14:353–63
51. Bond WF, Spillane L: The use of simulation for emergency medicine resident assessment. Acad Emerg Med 2002; 9:1295–9
52. Cantrell MJ, Deloney LA: Integration of standardized patients into simulation. Anesthesiol Clin 2007; 25:377–83
53. Nestel D, Kneebone R, Black S: Simulated patients and the development of procedural and operative skills. Med Teach 2006; 28:390–1
54. Issenberg SB, Scalese RJ: Simulation in health care education. Perspect Biol Med 2008; 51:31–46
55. Ogden PE, Cobbs LS, Howell MR, Sibbitt SJ, DiPette DJ:

Clinical simulation: Importance to the internal medicine educational mission. Am J Med 2007; 120:820–4

56. Carroll JD, Messenger JC: Medical simulation: The new tool for training and skill assessment. Perspect Biol Med 2008; 51:47–60
57. Gaba DM, DeAnda A: A comprehensive anesthesia simulation environment: Re-creating the operating room for research and training. ANESTHESIOLOGY 1988; 69:387–94
58. Murray DJ, Boulet JR, Avidan M, Krause KC, Henrichs B, Woodhouse J, Evers AS: Performance of residents and anesthesiologists in a simulation-based skill assessment. ANESTHESIOLOGY 2007; 107:705–13
59. Murray DJ, Boulet JR, Kras JF, Woodhouse JA, Cox T, McAllister JD: Acute care skills in anesthesia practice: A simulation-based resident performance assessment. ANESTHESIOLOGY 2004; 101:1084–95
60. Wilkinson TJ, Harris P: The transition out of medical school—A qualitative study of descriptions of borderline trainee interns. Med Educ 2002; 36:466–71
61. Murray D, Enarson C: Communication and teamwork: Essential to learn but difficult to measure. ANESTHESIOLOGY 2007; 106:895–6
62. Whelan GP, Boulet JR, McKinley DW, Norcini JJ, van Zanten M, Hambleton RK, Burdick WP, Peitzman SJ: Scoring standardized patient examinations: Lessons learned from the development and administration of the ECFMG Clinical Skills Assessment (CSA). Med Teach 2005; 27: 200–6
63. Adler MD, Trainor JL, Siddall VJ, McGaghie WC: Development and evaluation of high-fidelity simulation case scenarios for pediatric resident education. Ambul Pediatr 2007; 7:182–6
64. Scavone BM, Sproviero MT, McCarthy RJ, Wong CA, Sullivan JT, Siddall VJ, Wade LD: Development of an objective scoring system for measurement of resident performance on the human patient simulator. ANESTHESIOLOGY 2006; 105:260–6
65. Morgan PJ, Cleave-Hogg D, Guest CB: A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. Acad Med 2001; 76:1053–5
66. Gordon JA, Tancredi DN, Binder WD, Wilkerson WM, Shaffer DW: Assessment of a clinical performance evaluation tool for use in a simulator-based testing environment: A pilot study. Acad Med 2003; 78:S45–7
67. Morgan PJ, Lam-McCulloch J, Herold-McIlroy J, Tarshis J: Simulation performance checklist generation using the Delphi technique. Can J Anaesth 2007; 54:992–7
68. Wainer H. Estimating coefficients in linear models: It don't make no never mind. Psychol Bull 1976;83:213–7
69. Boulet JR, van Zanten M, De Champlain A, Hawkins RE, Peitzman SJ: Checklist content on a standardized patient assessment: An ex post facto review. Adv Health Sci Educ Theory Pract 2008; 13:59–69
70. Hunt EA, Walker AR, Shaffner DH, Miller MR, Pronovost PJ: Simulation of in-hospital pediatric medical emergencies and cardiopulmonary arrests: Highlighting the importance of the first 5 minutes. Pediatrics 2008; 121: e34–43
71. Girzadas DV Jr, Clay L, Caris J, Rzechula K, Harwood R: High fidelity simulation can discriminate between novice and experienced residents when assessing competency in patient care. Med Teach 2007; 29:452–6
72. Baker DP, Salas E, King H, Battles J, Barach P: The role of teamwork in the professional education of physicians: Current status and assessment recommendations. Jt Comm J Qual Patient Saf 2005; 31:185–202
73. van Zanten M, Boulet JR, McKinley DW, De Champlain A, Jobe AC: Assessing the communication and interpersonal skills of graduates of international medical schools as part of the United States Medical Licensing Exam (USMLE) Step 2 Clinical Skills (CS) Exam. Acad Med 2007; 82: S65–8
74. Weller JM, Bloch M, Young S, Maze M, Oyesola S, Wyner J, Dob D, Haire K, Durbridge J, Walker T, Newble D: Evaluation of high fidelity patient simulator in assessment of performance of anaesthetists. Br J Anaesth 2003; 90: 43–7
75. Boulet JR, Rebbecchi TA, Denton EC, McKinley DW, Whelan GP: Assessing the written communication skills of medical school graduates. Adv Health Sci Educ Theory Pract 2004; 9:47–60
76. Swanson DB, Clauser BE, Case SM: Clinical skills assessment with standardized patients in high-stakes tests: A framework for thinking about score precision, equating, and security. Adv Health Sci Educ Theory Pract 1999; 4:67–106
77. Holmboe ES, Cassel CK: The role of physicians and certification boards to improve quality. Am J Med Qual 2007; 22:18–25
78. Klass D: Assessing doctors at work—progress and challenges. N Engl J Med 2007; 356:414–5
79. Morgan PJ, Cleave-Hogg D, DeSousa S, Tarshis J: High-fidelity patient simulation: Validation of performance checklists. Br J Anaesth 2004; 92:388–92
80. Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R: Assessment of clinical performance during simulated crises using both technical and behavioral ratings. ANESTHESIOLOGY 1998; 89:8–18
81. Byrne AJ, Greaves JD: Assessment instruments used during anaesthetic simulation: Review of published studies. Br J Anaesth 2001; 86:445–50
82. Fletcher GCL, McGeorge P, Flin RH, Glavin RJ, Maran NJ: The role of non-technical skills in anaesthesia: A review of current literature. Br J Anaesth 2002; 88:418–29
83. Morgan PJ, Cleave-Hogg D, Desousa S, Lam-McCulloch J: Applying theory to practice in undergraduate education using high fidelity simulation. Med Teach 2006; 28:e10–5
84. van der Vleuten CP, Norman GR, De Graaff E: Pitfalls in the pursuit of objectivity: Issues of reliability. Med Educ 1991; 25:110–8
85. Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Noel AG, Szalai JP: Testing internal consistency and construct validity during evaluation of performance in a patient simulator. Anesth Analg 1998; 86:1160–4
86. Shayne P, Gallahue F, Rinnert S, Anderson CL, Hern G, Katz E: Reliability of a core competency checklist assessment in the emergency department: The Standardized Direct Observation Assessment Tool. Acad Emerg Med 2006; 13:727–32
87. Brett-Fleegler MB, Vinci RJ, Weiner DL, Harris SK, Shih MC, Kleinman ME: A simulator-based tool that assesses pediatric resident resuscitation competency. Pediatrics 2008; 121:e597–603
88. Boulet JR. Generalizability theory: Basics, Encyclopedia of Statistics in Behavioral Science. Edited by Everitt BS, Howell DC. Chichester, Wiley, 2005, pp 704–11
89. Thompson B. A brief introduction to generalizability theory, Score Reliability: Contemporary Thinking on Reliability Issues. Edited by Thompson B. Thousand Oaks, Sage Publications, Inc., 2003, pp 43–58
90. Vu NV, Barrows HS: Use of standardized patients in clinical assessments: Recent developments and measurement findings. Educ Res 1994; 23:23–30
91. Berkenstadt H, Ziv A, Gafni N, Sidi A: Incorporating simulation-based objective structured clinical examination into the Israeli National Board Examination in ANESTHESIOLOGY. Anesth Analg 2006; 102:853–8
92. Savoldelli GL, Naik VN, Joo HS, Houston PL, Graham M, Yee B, Hamstra SJ: Evaluation of patient simulator performance as an adjunct to the oral examination for senior anesthesia residents. ANESTHESIOLOGY 2006; 104:475–81
93. Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Murphy PM, Szalai JP: Testing the raters: Inter-rater reliability of standardized anaesthesia simulator performance. Can J Anaesth 1997; 44:924–8
94. Schwid HA, Rooke GA, Carline J, Steadman RH, Murray WB, Olympio M, Tarver S, Steckner K, Wetstone S: Eval-

uation of anesthesia residents using mannequin-based simulation: A multiinstitutional study. ANESTHESIOLOGY 2002; 97:1434–44

95. Ringsted C, Ostergaard D, Ravn L, Pedersen JA, Berlac PA, Van der Vleuten CPM: A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. Med Teach 2003; 25:654–8
96. Weller JM, Jolly B, Robinson B: Generalisability of behavioural skills in simulated anaesthetic emergencies. Anaesth Intensive Care 2008; 36:185–9
97. Weller JM, Robinson BJ, Jolly B, Watterson LM, Joseph M, Bajenov S, Haughton AJ, Larsen PD: Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. Anaesthesia 2005; 60:245–50
98. Murray DJ, Boulet JR, Kras JF, McAllister JD, Cox TE: A simulation-based acute skills performance assessment for anesthesia training. Anesth Analg 2005; 101:1127–34
99. Downing SM: Validity: On the meaningful interpretation of assessment data. Med Educ 2003; 37:830–7
100. Kane MT: Current concerns in validity theory. J Educ Meas 2001; 38:319–42
101. Clauser BE, Margolis MJ, Swanson DB: Issues of validity and reliability for assessments in medical education, Practical Guide to the Evaluation of Clinical Competence, 1st edition. Edited by Holmboe ES, Hawkins RE. Philadelphia, Mosby/Elsevier, 2008, pp 10–23
102. Grand'Maison P, Brailovsky CA, Lescop J: Content validity of the Quebec licensing examination (OSCE). Assessed by practising physicians. Can Fam Physician 1996; 42: 254–9
103. Berkenstadt H, Ziv A, Gafni N, Sidi A: The validation process of incorporating simulation-based accreditation into the ANESTHESIOLOGY Israeli national board exams. Isr Med Assoc J 2006; 8:728–33
104. Gordon JA, Wilkerson WM, Shaffer DW, Armstrong EG: "Practicing" medicine without risk: Students' and educators' responses to high-fidelity patient simulation. Acad Med 2001; 76:469–72
105. Farrell SE: Evaluation of student performance: Clinical and professional performance. Acad Emerg Med 2005; 12:302.e6–10
106. Rosenthal R, Gantert WA, Hamel C, Hahnloser D, Metzger J, Kocher T, Vogelbach P, Scheidegger D, Oertli D, Clavien PA: Assessment of construct validity of a virtual reality laparoscopy simulator. J Laparoendosc Adv Surg Tech A 2007; 17:407–13
107. Young JS, Stokes JB, Denlinger CE, Dubose JE: Proactive versus reactive: The effect of experience on performance in a critical care simulator. Am J Surg 2007; 193:100–4
108. Boulet JR, Murray D, Kras J, Woodhouse J: Setting performance standards for mannequin-based acute-care scenarios. Simul Healthc 2008; 3:72–81
109. Reznick RK, MacRae H: Teaching surgical skills—changes in the wind. N Engl J Med 2006; 355:2664–9
110. Tamblyn R, Abrahamowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, Smee S, Blackmore D, Winslade N, Girard N, Du Berger R, Bartman I, Buckeridge DL, Hanley JA: Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA 2007; 298:993–1001
111. Hatala R, Issenberg SB, Kassen B, Cole G, Bacchus CM, Scalese RJ: Assessing cardiac physical examination skills using simulation technology and real patients: A comparison study. Med Educ 2008; 42:628–36
112. Leape LL, Fromson JA: Problem doctors: Is there a system-level solution? Ann Intern Med 2006; 144:107–15
113. Fox-Robichaud AE, Nimmo GR: Education and simulation techniques for improving reliability of care. Curr Opin Crit Care 2007; 13:737–41
114. Knudson MM, Khaw L, Bullard MK, Dicker R, Cohen MJ, Staudenmayer K, Sadjadi J, Howard S, Gaba D, Krummel T: Trauma training in simulation: Translating skills from SIM time to real time. J Trauma 2008; 64:255–63
115. Ferris TG, Vogeli C, Marder J, Sennett CS, Campbell EG: Physician specialty societies and the development of physician performance measures. Health Aff (Millwood) 2007; 26:1712–9
116. Howard SK, Gaba DM, Fish KJ, Yang G, Sarnquist FH: Anesthesia crisis resource management training: Teaching anesthesiologists to handle critical incidents. Aviat Space Environ Med 1992; 63:763–70
117. Pian-Smith MCM, Simon R, Minehart RD, Podraza M, Rudolph J, Walzer T, Raemer D: Teaching residents the two-challenge rule: A simulation-based approach to improve education and patient safety. Simul Healthc 2009; 4:84–91
118. Fernandez R, Vozenilek JA, Hegarty CB, Motola I, Reznek M, Phrampus PE, Kozlowski SWJ: Developing expert medical teams: Toward an evidence-based approach. Acad Emerg Med 2008; 15:1025–36
119. Morgan PJ, Cleave-Hogg D: A Canadian simulation experience: Faculty and student opinions of a performance evaluation study. Br J Anaesth 2000; 85:779–81
120. Langdon MG, Cunningham AJ: High-fidelity simulation in post-graduate training and assessment: An Irish perspective. Ir J Med Sci 2007; 176:267–71
121. Yee B, Naik VN, Joo HS, Savoldelli GL, Chung DY, Houston PL, Karatzoglou BJ, Hamstra SJ: Nontechnical skills in anesthesia crisis management with repeated exposure to simulation-based education. ANESTHESIOLOGY 2005; 103: 241–8
122. Gaba DM, DeAnda A: The response of anesthesia trainees to simulated critical incidents. Anesth Analg 1989; 68: 444–51
123. Kuduvalli PM, Jervis A, Tighe SQM, Robin NM: Unanticipated difficult airway management in anaesthetised patients: A prospective study of the effect of mannequin training on management strategies and skill retention. Anaesthesia 2008; 63:364–9
124. Waldrop WB, Murray DJ, Boulet JR, Kras JF: Management of anesthesia equipment failure: A simulation-based resident skill assessment. Anesth Analg 2009; 109:426–33
125. Wright MC, Phillips-Bute BG, Petrusa ER, Griffin KL, Hobbs GW, Taekman JM: Assessing teamwork in medical education and practice: Relating behavioural teamwork ratings and clinical performance. Med Teach 2009; 31: 30–8
126. Morgan PJ, Cleave-Hogg DM, Guest CB, Herold J: Validity and reliability of undergraduate performance assessments in an anesthesia simulator. Can J Anaesth 2001; 48:225–33
127. Morgan PJ, Cleave-Hogg D: Evaluation of medical students' performance using the anaesthesia simulator. Med Educ 2000; 34:42–5
128. Weller JM, Merry AF, Robinson BJ, Warman GR, Janssen A: The impact of trained assistance on error rates in anaesthesia: A simulation-based randomised controlled trial. Anaesthesia 2009; 64:126–30
129. Domuracki KJ, Moule CJ, Owen H, Kostandoff G, Plummer JL: Learning on a simulator does transfer to clinical practice. Resuscitation 2009; 80:346–9
130. Kuduvalli PM, Parker CJR, Leuwer M, Guha A: Retention and transferability of team resource management skills in anaesthetic emergencies: The long-term impact of a high-fidelity simulation-based course. Eur J Anaesthesiol 2009; 26:17–22
131. Blum RH, Raemer DB, Carroll JS, Sunder N, Felstein DM, Cooper JB: Crisis resource management training for an anaesthesia faculty: A new approach to continuing education. Med Educ 2004; 38:45–55
132. Okuda Y, Bryson EO, DeMaria S Jr, Jacobson L, Quinones J, Shen B, Levine AI: The utility of simulation in Medical education: What is the evidence? Mt Sinai J Med 2009; 76:330–43